Assessing school effects without controlling for prior achievement?

Hans Luyten*, Peter Tymms & Paul Jones****

* *University of Twente, The Netherlands*

** *CEM Centre, Durham University, United Kingdom*

Abstract

The research findings presented in this paper illustrate how the "value added" of schooling can be assessed empirically using cross-sectional data. Application of the regression-discontinuity approach within a multilevel framework produces both an estimate of the absolute effect of one year schooling and an estimate of the variation across schools of this effect. In the study reported here the approach was applied to both a cross-sectional and a longitudinal dataset. The research findings indicate to what extent different results are produced when cross-sectional as opposed to longitudinal data are analysed.

Keywords: School effectiveness, Regression discontinuity, Multilevel analysis, Primary Education, PIPS

Correspondence:

Hans Luyten

Faculty of Behavioural Sciences

Department of Educational Organisation and Management

University of Twente

P.O. Box 217

7500 AE Enschede

The Netherlands

E-mail: j.w.luyten@gw.utwente.nl

Assessing school effects without controlling for prior achievement?

Introduction

Most researchers in the field of school effectiveness would agree that for a valid assessment of the “value added” by schools it is highly desirable to take each pupil’s prior achievement into account. The multilevel analyses that are usually conducted focus on differences in achievement between schools. Adjusting for the bias that results from intake differences seems essential. By far the most frequently applied method includes prior achievement (usually together with other background characteristics) as a covariate in the data analysis. In this paper we point to an alternative method that does not need controls for prior achievement, namely the regression-discontinuity approach. Our data analysis will indicate to what extent this method produces different outcomes when a cross-sectional as opposed to a longitudinal dataset is analyzed.

The regression-discontinuity approach is a useful method for assessing the absolute effect of schooling with cross-sectional data. Studies based on this approach indicate that more than 50% of the progress pupils make over a one-year period is accounted for by schooling. (Cahan & Davis, 1987; Luyten, 2006). This perspective differs considerably from the results presented in the bulk of the school effectiveness studies, which typically report that schools account for approximately ten percent of the variance in test scores after controlling for prior attainment (Scheerens & Bosker, 1997). The figure of 50% refers to the impact of receiving education in the upper grade as opposed to the lower grade and is calculated as a percentage change in test score, whereas the figure of 10% refers to the variation in the impact of schools[1]. The 50% figure is more in line with educational effectiveness studies that address the variation between both schools and classes/teachers (e.g. Hill & Rowe, 1996; Opdenakker & Van Damme, 2000).

Regression-discontinuity capitalizes on the fact that students are assigned to a higher or lower grade on the basis of their date of birth. In most countries pupils born before a certain cut-off point (in England, 1st September) generally end up in a higher grade than pupils that are just a few days younger. The effect of one year extra schooling can thus be assessed by comparing the achievement scores of the pupils born shortly before the cut-off point to the scores of the ones

[1] Note that these percentages are not really comparable with one another. Both figures express different aspects of the same phenomenon. It seems possible to convert both percentages to effect sizes that have been defined in relation to interventions in which there is a control and an experimental group (see Tymms, 2004). We will briefly return to this matter in the final section, but a detailed coverage of this topic is beyond the scope of this article.

born shortly after the cut-off point. The latter ones are in the lower grade, whereas the former are in the upper grade and have received an extra year of schooling. Within each grade the relationship between age and achievement is estimated. If the data analysis reveals a discontinuity between the oldest pupils in the lower grade and the youngest ones in the upper grade, this is interpreted as the effect of one year extra schooling (i.e., being in the upper grade). As the analysis takes into account the impact of the criterion used for assigning students to either grade, alternative interpretations are largely ruled out. It is conceivable that the cut-off point coincides with other relevant factors, but – generally speaking – this is unlikely. The main strength of the regression-discontinuity approach is that it predicts an effect at a very specific point on the age continuum. When this approach is applied within a multilevel framework, the effect of schooling can be modelled as a random effect at the school level. Thus differences in effectiveness between schools can also be estimated.

Note, however, that correct modelling of the relation between age and achievement is crucial. If a linear function is estimated, while in reality the function is quadratic or cubic the model will be misspecified, and the regression-discontinuity may be biased. Another essential requirement is adherence to the cut-off point. In many countries, assignment to a particular grade does not adhere fully to a nation-wide cut-off point. In some cases, this may be (partly) due to regional variations (for example, Australia and the United States), but grade retention is usually the main cause. If the degree of "misclassification" is not excessive, it is still possible to obtain reliable effect estimates. If the percentage of misclassified participants is limited (preferably below five percent), it is best to exclude them (Judd & Kennedy, 1981; Trochim, 1984; Shadish et al., 2002). In the dataset analyzed for the present study less than 1.5% of the pupils were in the "wrong" grade given their date of birth.

One specific advantage of the regression-discontinuity approach is that – in principle – it does not require controlling for prior achievement (or any other background characteristics) in order to assess the effect of schooling. As in randomised experiments, cross-sectional data is sufficient for strong causal inferences. The basic assumption is that the selection criterion constitutes the only relevant difference between control and treatment group. With regard to pupils from two (or more) consecutive grades this means that the distinct groups are assumed to be equivalent in all relevant respects (e.g. gender distribution, socioeconomic background, cognitive aptitudes) besides age. This implies that estimating the grade effect by means of longitudinal data should yield results that hardly differ from an analysis of cross-sectional data.

This assertion will be put to the test in the present study. The effect of schooling will be estimated using both longitudinal and cross-sectional data. The data relate to four and five year old pupils in English reception classes. The reception class appears to be unique to England. In other countries it is more likely to be identified as a kindergarten group but in England it is clearly part of the school and typically follows pre-school provision (nursery or play group).

Modelling the absolute effect of schooling

Equation (1) presents the basic regression-discontinuity model. The coefficients β_1 and β_2 express the effect of age and the effect of one extra year schooling. It applies to a dataset with students from two consecutive grades. The effect of age is assumed linear and identical in both grades (i.e., no interaction between age and grade).

$$Y_i = \beta_0 + \beta_1(x_i - x_0) + \beta_2 z_i + R_i \qquad (1)$$

Where:

Y_i = *outcome measure (e.g., mathematics achievement score for student i)*
x_i = *age, pupil i (see table 2 for details on the coding of the age variable)*
x_0 = *cut-off value (in this study: 5 years)*
z_i = *grade, student i (0 if lower grade; 1 if upper grade)*
β_0 = *parameter for comparison group intercept at cut-off*
β_1 = *age effect*
β_2 = *effect of being in the upper grade (i.e., having received an extra year of schooling)*
R_i = *random residual*

When combining the regression-discontinuity approach with multilevel analysis, the intercept (β_0) and the effect of one extra year of schooling (β_2) are allowed to vary across schools:

$$Y_{ij} = \beta_{0j} + \beta_1(x_{ij} - x_0) + \beta_{2j} z_{ij} + R_{ij} \qquad (2)$$

j = *index for schools*
i = *index for students within schools*

The intercept and the effect of schooling are now school-dependent. These school-dependent coefficients can be separated into an average coefficient and the school-dependent deviation:

$$\beta_{0j} = \gamma_{00} + U_{0j}$$
$$\beta_{2j} = \gamma_{20} + U_{2j}$$

Substitution leads to the following model:

$$Y_{ij} = \gamma_{00} + \beta_1(x_{ij} - x_0) + \gamma_{20}z_{ij} + U_{0j} + U_{2j}z_{ij} + R_{ij} \qquad (3)$$

In this equation, γ_{20} expresses the general effect of one year extra schooling, while U_{2j} represents the school-dependent deviation. Its variance is of particular interest, as it expresses the extent to which the effect of schooling differs between schools. When fitting this model, the variances of R_{ij} and U_{0j} are estimated, as is the covariance between U_{0j} and U_{2j}. This basic model can be extended in several ways. It can be applied to a wider range of adjacent grades (see the contribution by Kyriakides & Luyten to this special issue for an example) and the effect of age may be allowed to vary between grades. Such an interaction between the selection variable (age) and the independent variable of interest (grade), however, may simply reflect a curvilinear relation (Shadish et al., 2002; Trochim, 1984). In the findings section we will report the estimated effects of grade when the age-achievement relation is modelled as a quadratic function[2]. It is also possible to test whether the effect of schooling is dependent on other variables (e.g., SES). This could be modelled through interaction effects between the additional variable(s) and the grade effect. A significant interaction effect would imply that the effect of schooling varies for pupils with different socioeconomic backgrounds. This question will not be addressed in the present contribution. The contribution by Verachtert et al. to this issue pays explicit attention to differences in learning rates between pupils from different backgrounds. Another possible extension would be to model a random effect of age at the school level. Fitting complex models with multiple random effects, however, requires a (very) large dataset. The dataset analyzed for this study includes data from only 18 schools. As a result only relatively simple models could be fitted.

[2] The estimated effects of grade in the models with quadratic effects of age included (see table 3) indeed hardly differ from the effects of grade in models that include an interaction effect of age with grade. Details on these analyses can be obtained from the first author.

Straightforward application of the regression-discontinuity approach relates to cross-sectional data. The achievement scores of pupils in the upper grade are then compared to the scores of other pupils, which are in the lower grade. In the present study, however, the approach is also applied to longitudinal data. In that case the achievement scores of the pupils in the upper grade are compared to the scores attained one year earlier by the same pupils. In both cases we take the effect of age on achievement into account. The net effect of being in the upper grade is the difference in achievement between the lower and the upper grade adjusted for the effect of age. In the case of longitudinal data the regression discontinuity analysis uses matched pupils whereas the cross-sectional approach uses different pupils. Analysis of the longitudinal data yields additional information that cannot be obtained with cross-sectional data, namely the variance of the grade effect at the individual level. It seems plausible that the effect of having received an extra year of schooling varies not only across schools, but also between pupils within schools. Estimation of this variance, however, requires longitudinal data.

Research questions

The primary aim of this study is to compare research outcomes based on cross-sectional data with outcomes that are based on longitudinal data. More specifically this is applied to the following research questions:

1. What is the effect of one extra year of schooling on the level achievement?
2. What proportion of the difference in achievement between two grades is accounted for by schooling?
3. To what extent does the effect of one extra year of schooling vary between schools?

The data analyses relate to three subject areas, namely mathematics, reading and phonics. Some more details regarding the outcome measures are provided in the variables section.

Data

The data come from the Performance Indicators in Primary School (PIPS) project (Tymms, 1999a; Tymms & Albone, 2002) which is run from the Curriculum Evaluation and Management (CEM) Centre (Tymms & Coe, 2003) in Durham University, England. The project is designed to monitor the progress of pupils in school for schools. The data are not intended for outsiders and the schools, or their authorities, pay to join the project or part of it. The PIPS project starts with an on-entry baseline assessment within six weeks of the pupils' start at school. It is repeated at the end of the academic year in an extended form. The assessment can be administered using a paper based or a computerised form and is always administered in a one-to-one situation taking about twenty minutes per pupil.

Our analyses relate to a (very) small part of the entire PIPS dataset, namely to the schools where the pupils were first assessed using the on-entry baseline and re-assessed in September of the following year. In most schools the pupils are assessed at the start and the end of the school year (September and June respectively). Only 18 schools are thus included in the analyses. The analyses focus exclusively on pupils with standard school careers. Those with delayed or accelerated careers (1.5%) were excluded from the analyses.

The analyses focus on two datasets, namely a cross-sectional and a longitudinal one. The pupils in the upper grade (called "Year 1" in England) are the same in both datasets, but the lower grade (reception) pupils are not. The analyses of the cross-sectional data relate to a comparison of two distinct groups of pupils that were assessed at the same time, namely in September 2004. The upper grade pupils were born in the period September 1998 – August 1999 and the lower grade ones one year later, in the period September 1999 – August 2000. The longitudinal analyses relate to a comparison of the scores obtained by the same pupils at two different points in time, namely in September 2004 (when they were in the upper grade) and September 2003 (when they were in the lower grade and thus one year younger). See also table 2.

Variables

The dependent variables in the analysis are pupil achievement for mathematics, reading and phonics. The test was constructed to provide measures of the best predictors of later achievement in school. As such it includes the precursors to reading and mathematics and this extends to reading *per se* and into mathematics (arithmetic) (see Tymms, 1999b). There are 17 sections in the assessment with some additional voluntary sections, which are not used in this paper. The sections can be grouped into reading, mathematics and phonics. Areas assessed include name writing, picture vocabulary, concepts about print, letter identification, reading simple words and reading more complex sentences, which combine into the reading score; word repeating and identifying pairs of rhyming words, which combine into the phonics score; ideas about maths, counting, informal simple sums, number recognition, shapes and formal maths which combine into the mathematics score. An example will give a feel for the test.

For letter identification the child is shown letters of the alphabet in order of their empirically established difficulty. It starts with the first letter of the child's name and proceeds through many letters with a mixture of upper and lower case letters. The test algorithm is that when three in a row are answered incorrectly or four altogether the section is halted and the child moves to the next section. Some more sample questions are presented at the home page of the CEM centre (http://www.cemcentre.org/).

Schools use the same assessment both on entry to school and for the follow-up assessment. In order to reduce the time spent by schools on the re-test, pupils are not re-assessed on sections they completed adequately at the start of the year and instead are credited with full marks in those sections. However, whilst this saves a lot of time for schools, there is the risk that some children will be credited for items that they would have got wrong. We are able to empirically measure the size of this difference using data from schools where something unfortunate has happened to their computers. If the computers are stolen, broken or even upgraded, if the school did not keep a backup copy of the data, the children will have to complete the final assessment from the very beginning. This means they are assessed on those sections for which they should have been credited full marks. As such we can compare their score on the full test with the calculated score if they had been credited the marks. On average this makes a difference of 2.77 items and this correction was applied before the models were set up.

The assessment has a test-retest reliability of 0.98 (CEM website). As stated previously, the assessment was designed to predict later attainment. The correlations between the assessment scores at the start of the year, the end of the year, and later PIPS assessments up to the end of primary school, six years later, are all around 0.7 (Tymms 1999, Tymms et al, 2000, and Tymms et al, 2007).

Table 1 shows the mean and standard deviation for each subject by grade. Figures 1a, 1b and 1c provide a graphical display of the frequency distributions of the scores for maths, reading and phonics. Figure 1a shows that in both lower and upper grade the distributions of the maths scores present fairly close approximations of the standard normal distribution. The pupils in the upper grade get higher scores, but the amount of variation hardly differs between both grades. Moreover, the distributions for the cross-sectional and longitudinal data in the lower grade are highly similar. Most important is the apparent absence of any floor or ceiling effects in the distributions, as these might bias the effect estimates. Figure 1b shows the frequency distributions for reading. Also in this case the distributions show a fair degree of resemblance with the standard normal distribution in both grades, but the graph displays much more variation for reading in the upper grade. This is also expressed in the size of the standard deviations (see table 1). There is little evidence for any floor or ceiling effects. In the lower grade, the longitudinal and cross-sectional data produce similarly shaped distributions. Figure 1c, however, shows an unmistakable ceiling effect in the distribution of the phonics scores in the upper grade. In fact 46% of the pupils in the upper grade attained the maximum score on the phonics test. This implies that many pupils would most probably have got a higher score, if the discriminatory power of the test had been higher. However, the ceiling effect is not so surprising when one considers the limited scope of the test (word repeating and identifying pairs of rhyming words). With regard to our basic research question, the most important point is that the bias relates to both the cross-sectional and longitudinal dataset. Only if a floor or ceiling effect occurs exclusively in either dataset is it possible that this leads to different findings for the cross-sectional versus the longitudinal data. The distributions in the lower grade are again quite similar for both the cross-sectional and longitudinal data.

Table 1: Basic statistics given years for when the data were collected

	LOWER GRADE		**UPPER GRADE**
	cross-sectional data	**longitudinal data**	**cross-sectional and longitudinal**
Mathematics			
Average score	24.50	24.67	40.42
Standard deviation	9.14	8.58	8.43
Reading			
Average score	30.45	29.42	88.51
Standard deviation	19.23	15.54	36.00
Phonics			
Average score	9.07	9.11	14.57
Standard deviation	4.46	4.60	3.40
Number of pupils	593	599	599

Grade and age are the independent variables in the analyses. Grade was recoded to assign scores of zero to students in the lower grade and scores of one to students in the upper grade. The variable that denotes a pupil's age at the date of assessment is based on year and month of birth. Each age was transformed into a decimal number. For example, a pupil born in March 1999 and assessed in September 2004 received a score of 5.50, and a pupil born in April received a score of 5.42. The cut-off value (5.00) was then subtracted from these scores, giving each of the oldest pupils in the lower grade (the comparison group) a score of zero. Table 2 illustrates the transformation of the original birth dates to the scores used in the analyses.

Figure 1a

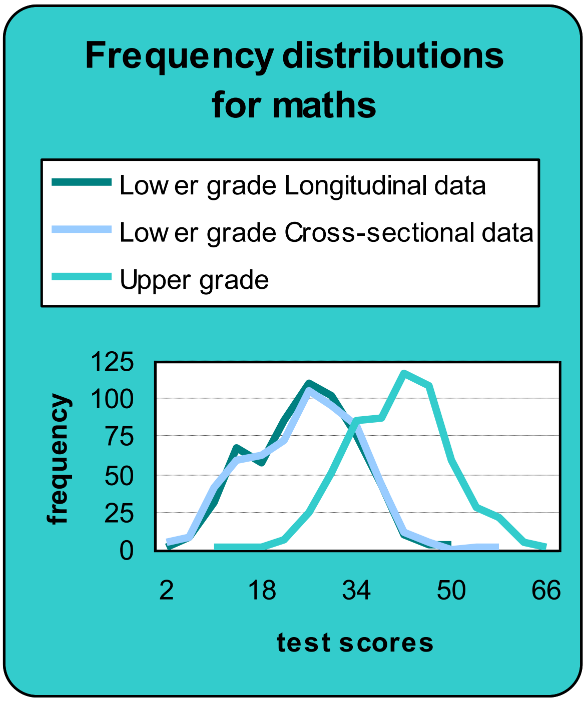


Figure 1b

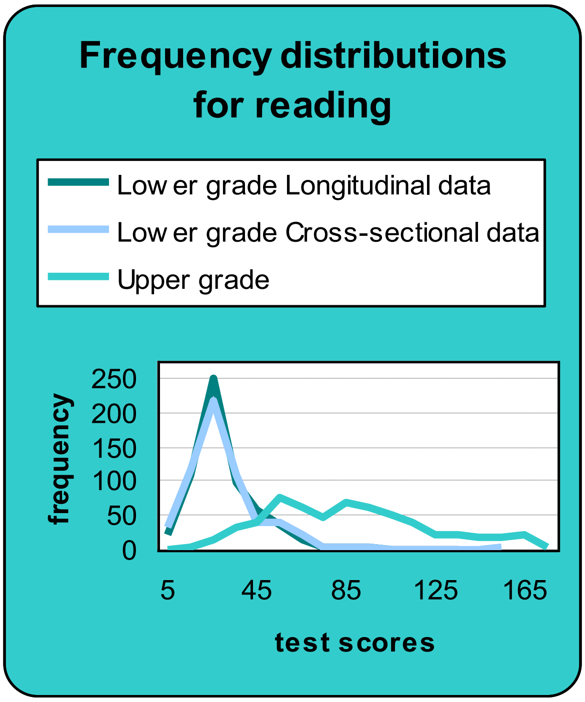


Figure 1c

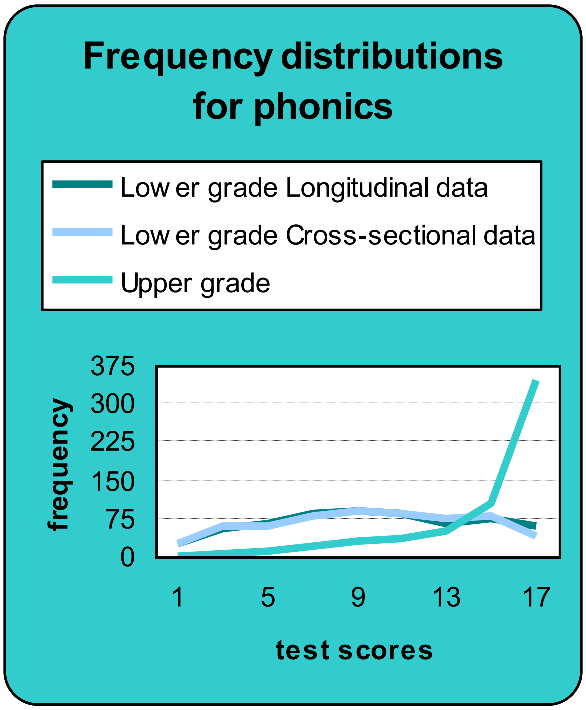


Table 2: Assessment dates, dates of birth and ages at date of assessment (cutoff = 5.00)

	CROSS-SECTIONAL DATA				LONGITUDINAL DATA			
		date of birth	**age in decimals**	**minus cut-off**		**date of birth**	**age in decimals**	**minus cut-off**
Upper grade	**Assessment date September 2004**	Sept. 1998	6.00	1.00	**Assessment date September 2004**	Sept. 1998	6.00	1.00
		Oct. 1998	5.92	.92		Oct. 1998	5.92	.92
		Nov. 1998	5.83	.83		Nov. 1998	5.83	.83
		Dec. 1998	5.75	.75		Dec. 1998	5.75	.75
		Jan. 1999	5.67	.67		Jan. 1999	5.67	.67
		Feb. 1999	5.58	.58		Feb. 1999	5.58	.58
		March 1999	5.50	.50		March 1999	5.50	.50
		April 1999	5.42	.42		April 1999	5.42	.42
		May 1999	5.33	.33		May 1999	5.33	.33
		June 1999	5.25	.25		June 1999	5.25	.25
		July 1999	5.17	.17		July 1999	5.17	.17
		Aug. 1999	5.08	.08		Aug. 1999	5.08	.08
Lower grade	**Assessment date September 2004**	Sept. 1999	5.00	.00	**Assessment date September 2003**	Sept. 1998	5.00	.00
		Oct. 1999	4.92	-.08		Oct. 1998	4.92	-.08
		Nov. 1999	4.83	-.17		Nov. 1998	4.83	-.17
		Dec. 1999	4.75	-.25		Dec. 1998	4.75	-.25
		Jan. 2000	4.67	-.33		Jan. 1999	4.67	-.33
		Feb. 2000	4.58	-.42		Feb. 1999	4.58	-.42
		March 2000	4.50	-.50		March 1999	4.50	-.50
		April 2000	4.42	-.58		April 1999	4.42	-.58
		May 2000	4.33	-.67		May 1999	4.33	-.67
		June 2000	4.25	-.75		June 1999	4.25	-.75
		July 2000	4.17	-.83		July 1999	4.17	-.83
		Aug. 2000	4.08	-.92		Aug. 1999	4.08	-.92

Models fitted

The first model fitted serves as the reference basis and relates only to the difference in achievement between lower and upper grades and to the variation of this difference between schools for each of the three subject areas in both the cross-sectional and the longitudinal dataset. The second model includes the effect of the pupil's age, thereby representing the basic regression-discontinuity approach within a multilevel framework, as described in equation (3). In this equation the relation between age and achievement is modelled as a linear function. Additional models were fitted to explore the empirical validity of curvilinear relationships between age and achievement. We only report the findings that include a quadratic term as the analyses that included cubic terms failed to produce significant improvements of the model fit. The analyses of the longitudinal data entail one important addition. For these data the variance of the effect of one year schooling is estimated both at the school and pupil level. For the cross-sectional data the variance of this effect can only be estimated at the school level, as the difference between the scores in the lower and upper grade cannot be measured at the individual level. The MLwiN software (Rasbash et al., 2000) was used to analyse the data.

Findings

Table 3 shows the basic findings of the analyses. The first model estimates the gross effect of being in the upper grade. The second model includes the linear effect of age. In the third model a quadratic term is included. The random effects in table 3 relate to this model (the random effects for models 1 & 2 are reported in appendix 1). For all three subjects, in all three models and for both the cross-sectional and longitudinal data the effect of being in the upper grade and its variance is found to be statistically significant ($\alpha < .05$; two-tailed) given the size of the relevant standard errors.

The gross effect of one year schooling, which is estimated in model 1, is very similar in the cross-sectional and longitudinal dataset for all three “subjects”. Models 2 and 3 show larger differences between both datasets with regard to the fixed effects of grade, but when these differences are compared to the standard errors of the effects, none of them can be considered statistically significant. Note that for mathematics and phonics the sign of the quadratic effect is negative in both the longitudinal and cross-sectional datasets, whereas for reading the sign is positive in both datasets. A negative sign of the quadratic term indicates that the positive effect of age on achievement decreases, as pupils grow older. A positive sign suggests the opposite. The figures 2a, 2b and 2c present graphical displays of the age achievement relationships and the discontinuities between the oldest pupils in the lower grade and youngest in the upper grade. The figures relate to the effects estimated in model 3, which includes a quadratic term for the effect of age. For mathematics and reading there is hardly any difference between the lines that represent the age achievement relation for the cross-sectional and the longitudinal data. Only for phonics, which is the most limited in scope of the three measures, is the difference easier to discern.

Table 3: Comparison of findings with longitudinal and cross-sectional data

	MATHEMATICS				READING				PHONICS			
	Cross-sectional data		Longitudinal data		Cross-sectional data		Longitudinal data		Cross-sectional data		Longitudinal data	
FIXED EFFECTS												
Model 1												
Intercept	24.63	(.84)	24.85	(.93)	31.25	(1.79)	31.35	(2.52)	9.43	(.52)	9.46	(.58)
Grade	15.81	(1.22)	15.58	(.91)	57.85	(3.50)	57.91	(2.72)	5.35	(.49)	5.33	(.42)
Model 2												
Intercept	28.59	(.94)	28.34	(.99)	40.02	(2.14)	38.13	(2.63)	10.50	(.55)	10.61	(.62)
Grade	7.21	(1.48)	7.75	(1.35)	38.77	(4.01)	42.80	(3.30)	3.03	(.61)	2.75	(.62)
Age	8.43	(.78)	7.83	(1.00)	18.71	(2.25)	15.07	(1.86)	2.26	(.36)	2.56	(.45)
Model 3												
Intercept	29.06	(1.00)	28.78	(1.00)	39.07	(2.11)	40.70	(2.68)	10.56	(.58)	11.11	(.65)
Grade	7.08	(1.48)	7.55	(1.35)	32.48	(4.16)	30.19	(4.00)	3.01	(.61)	2.31	(.64)
Age	8.65	(.80)	8.15	(1.01)	23.84	(2.53)	26.63	(2.91)	2.30	(.37)	3.10	(.48)
Age^2	-1.20	(.76)	-1.07	(.43)	11.06	(2.56)	9.15	(1.78)	-.14	(.36)	-.91	(.25)
RANDOM EFFECTS (Model 3)												
School level												
Variance intercept	11.08	(4.54)	11.01	(4.53)	47.08	(19.71)	103.05	(37.40)	4.36	(1.66)	5.55	(2.11)
Variance grade effect	22.71	(9.28)	12.91	(4.87)	122.79	(58.62)	89.16	(38.67)	3.17	(1.39)	2.85	(1.13)
Covariance	-7.88	(5.21)	-3.43	(3.48)	7.11	(24.16)	-6.22	(26.99)	-3.33	(1.42)	-3.73	(1.47)
Correlation	-.50 (not sign.)		-.29 (not sign.)		.09 (not sign.)		-.06 (not sign.)		-.90		-.94	
Pupil level												
Variance intercept	66.82	(3.94)	59.52	(3.49)	316.16	(18.64)	179.96	(10.56)	15.43	(.91)	16.52	(.97)
Variance grade effect	---	---	34.70	(2.04)	---	---	634.32	(37.19)	---	---	12.28	(.72)
Covariance	-7.35	(2.49)	-21.02	(2.08)	359.78	(31.76)	111.27	(14.75)	-2.52	(.55)	-9.21	(.70)
Correlation	cannot be computed		-.46		cannot be computed		.33		cannot be computed		-.65	

The figures in between brackets denote the standard errors.
The random effects for models 1and 2 are reported in appendix 1.

Table 3 also shows that the variance of the grade effect between schools is larger in the cross-sectional dataset for all three subjects. The difference between both datasets for reading equals 33.63 (122.79 – 89.16), which is smaller than the standard error in either dataset (namely 58.62 and 38.67). The same goes for phonics: the difference amounts to .32 (3.17 – 2.85), while the standard errors equal 1.39 and 1.13. With regard to mathematics the findings are less clear-cut. The difference between both datasets equals 9.80 (22.71 – 12.91), which is hardly larger than the standard error in the cross-sectional dataset (9.28) but over twice as large as the standard error in the longitudinal dataset (4.87). With regard to the basic research question of this study the most important finding is that the cross-sectional data do not produce an outcome that contradicts the estimate based on the longitudinal data. If one would use the cross-sectional data to construct a 95% confidence interval for the school level variance of the grade effect, it would still include the value obtained with the longitudinal data. It should be noted, though, that the low number of schools in this study (18) is bound to produce large standard errors and wide confidence intervals. It is also worth mentioning that when the variation of the grade effect across schools is expressed in terms of standard deviations rather than variances, the disparities between the cross-sectional and the longitudinal results look much more modest. In that case the result for mathematics is 4.77 ($\sqrt{22.71}$) in the cross-sectional dataset versus 3.59 ($\sqrt{12.91}$) in the longitudinal dataset. For reading the contrast is 11.08($\sqrt{122.79}$) vs. 9.44($\sqrt{89.16}$) and for phonics it is 1.78($\sqrt{3.17}$) vs. 1.69($\sqrt{2.85}$).

The findings for reading show some differences between the longitudinal and cross-sectional data with regard to the intercept variance at both the school and pupil level. The school level variance is considerably larger in the longitudinal dataset, while the pupil level variance is much lower. This means that with regard to the school and pupil level variance for reading the pupils in the cross-sectional dataset are not equivalent to the ones in the longitudinal dataset. Note that both datasets display much more similarity with regard to the variances of the grade effect for reading between schools. Even though the cross-sectional and longitudinal data show some differences with regard to the variation in reading scores (see also table 2), this does not preclude fairly similar outcomes concerning the variance of the grade effect.

Figure 2a

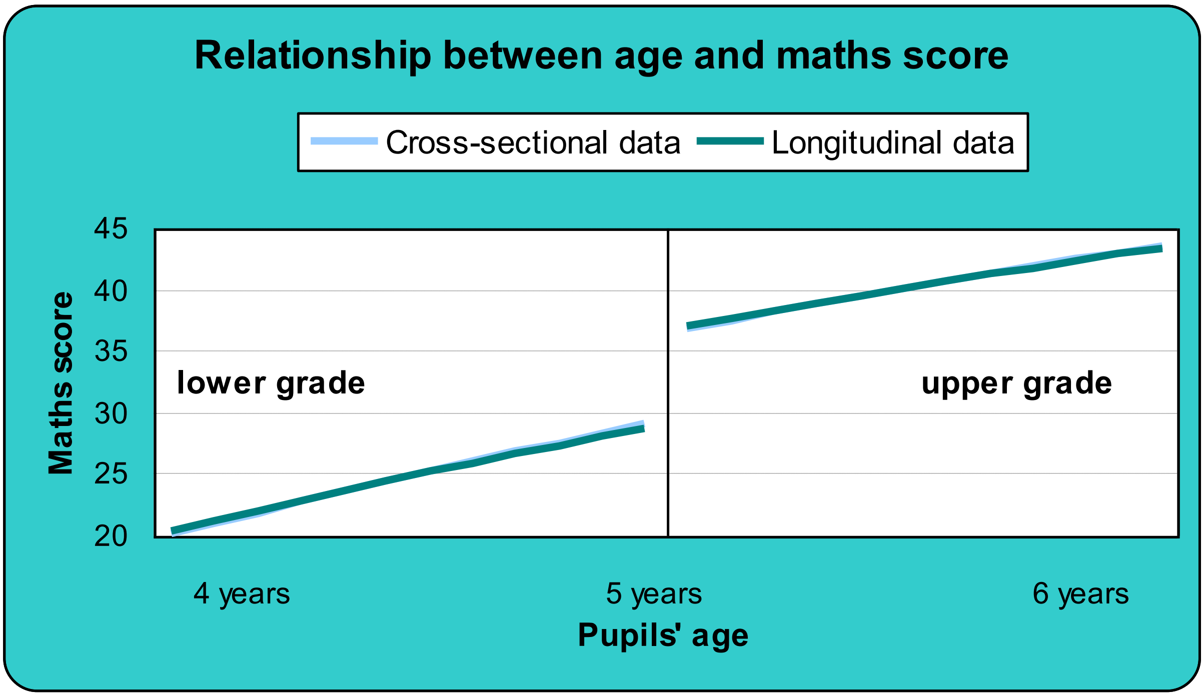


Figure 2b

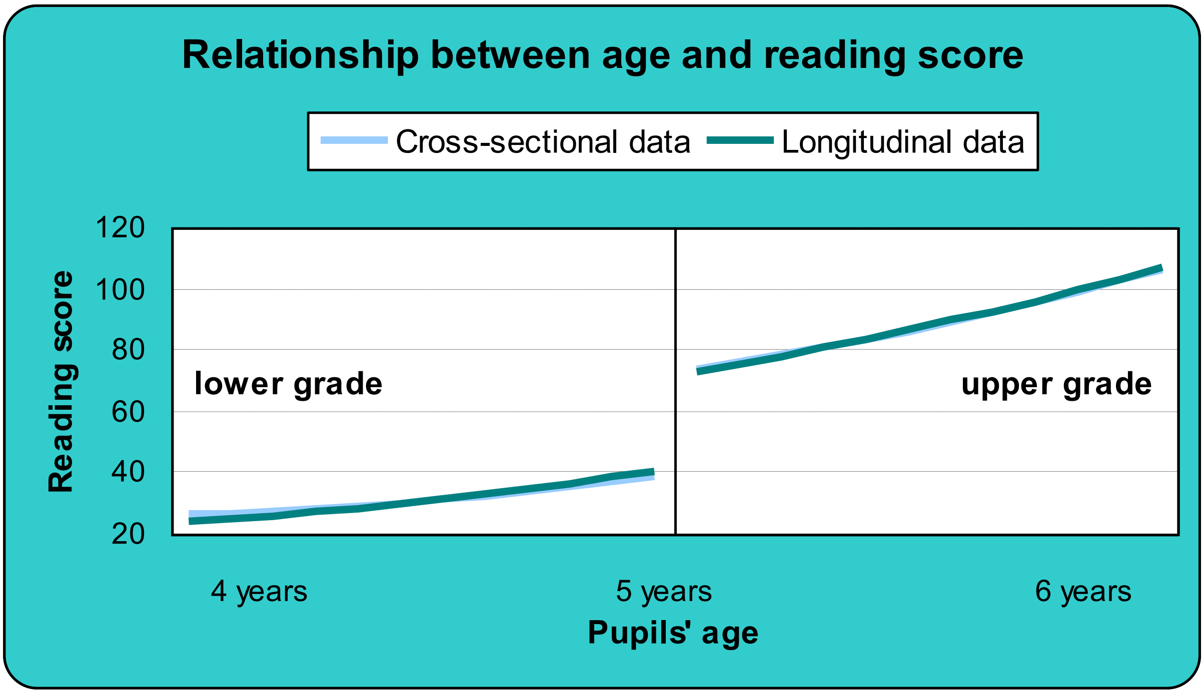
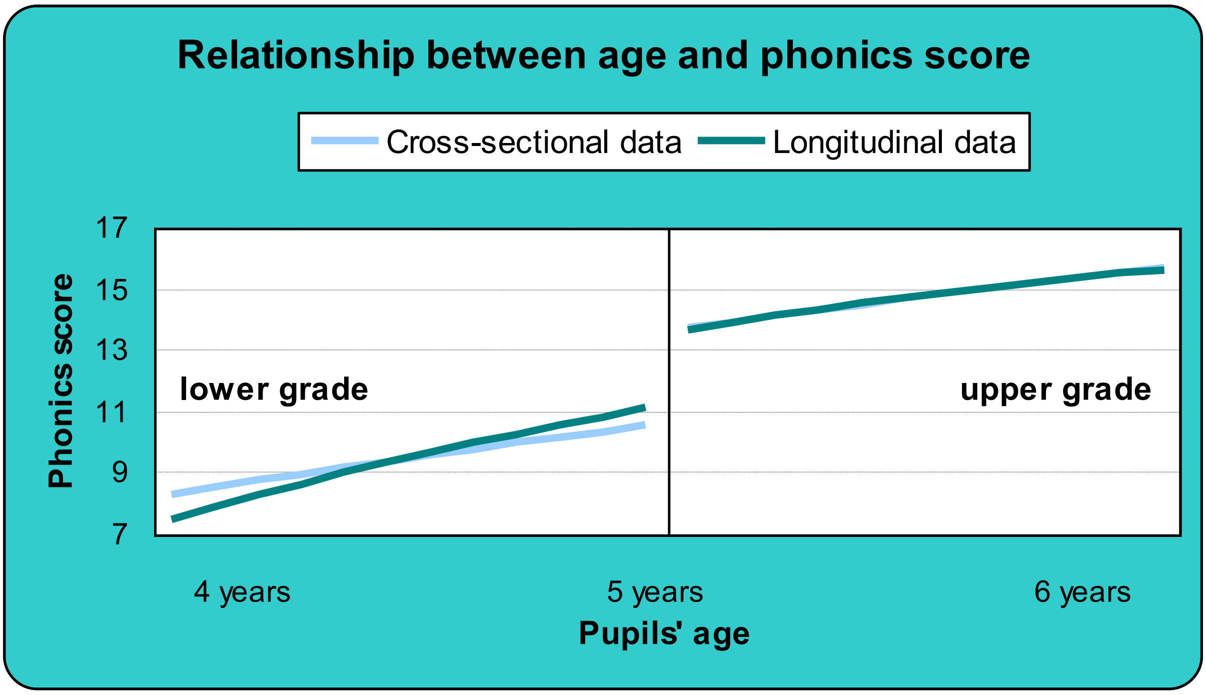


Figure 2c



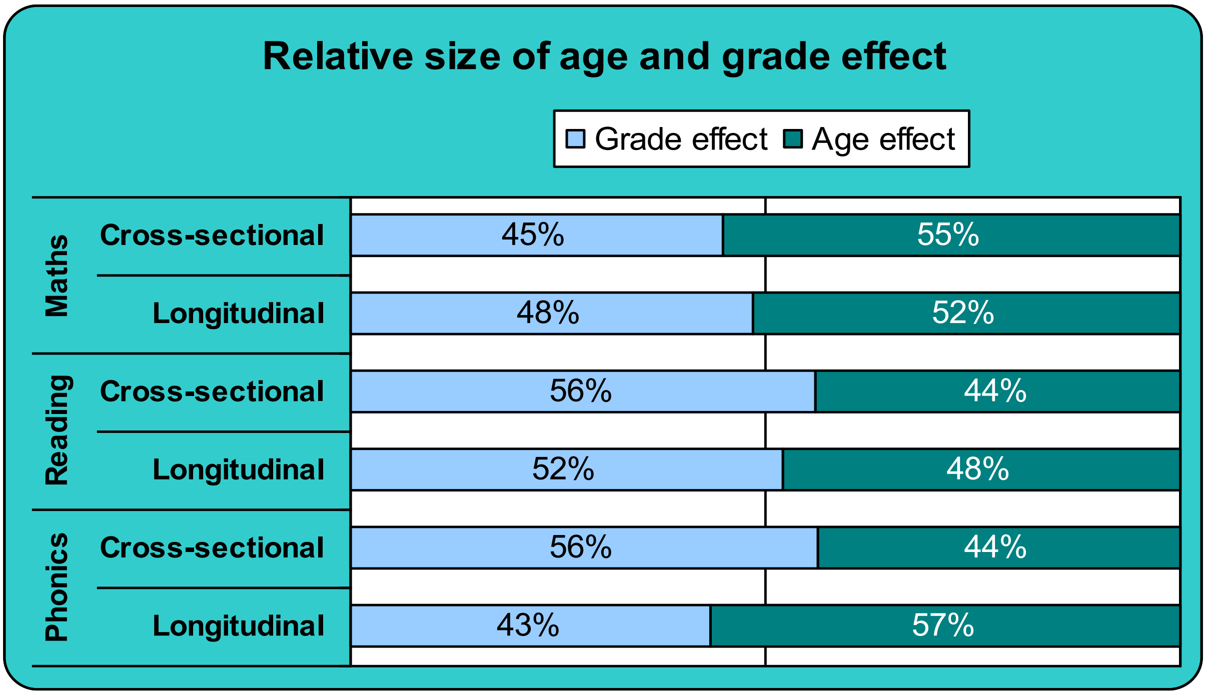
With regard to mathematics and reading none of the covariances at the school level between the intercept and the grade effect are significantly different from zero in either dataset. In the case of mathematics this is largely due to the low number of schools. The actual sizes of the covariances are substantial. The corresponding correlations equal -.50 and -.29 for the cross-sectional and longitudinal data respectively. For phonics, the covariance does deviate significantly from zero in both datasets. The negative covariance indicate that the effect of one year schooling is relatively low in schools with high intercepts, i.e. in schools where the scores for phonics are high in the lower grade. The covariances correspond to correlations equal to -.90 and -.94 in the cross-sectional and longitudinal dataset respectively. The main cause for these extremely strong correlations is presumably the ceiling effect in the frequency distribution of the phonics scores in the upper grade (see figure 1c). This precludes substantial progress in schools with high levels of achievement in the lower grade.

For the longitudinal datasets it also possible to calculate the correlation between the intercept and the grade effect at the pupil level. For the cross-sectional data the variance of the effect of one year schooling can only be estimated at the school level. The covariances at the pupil level for the cross-sectional data only express the extent to which the pupil level variance is higher or lower in the upper grade. The negative covariances for mathematics and phonics indicate less pupil level variance in the upper grade and the positive covariance for reading denotes more variance. All pupil level covariances are statistically significant. The covariance for phonics points to a strong and negative correlation (-.65) between intercept and grade effect at the pupil level. The effect of one year extra schooling is relatively small for pupils with high scores for phonics in the lower grade. It seems plausible that this (partly) reflects the ceiling effect in the distribution of the phonics scores (see figure 1c). A negative correlation (-.46) was also found for mathematics. The correlation for reading is positive (.33), which implies that the grade effect for reading is stronger if a pupil scored already high on reading in the lower grade.

The effect of one year schooling for mathematics in model 3 is 7.08 and 7.55 in the cross-sectional and longitudinal dataset respectively. For reading the effects are 32.48 and 30.19 and for phonics they are 3.01 and 2.31. These figures reflect the difference in achievement between the upper and lower grade after adjusting for the effect of age. When comparing the adjusted differences to the unadjusted differences (see table 3, model 1) one can compute the proportion of the difference in achievement between two grades that is accounted for by schooling. For example, the unadjusted difference between the lower and upper grade is 15.81 for mathematics in the cross-sectional dataset and the adjusted difference is 7.08. About 45% of the original difference remains after adjusting for age. Therefore 55% of the difference between grades must be attributed to the effect of age and 45% to the effect of schooling. Figure 3 displays the percentages for all three subjects in both the cross-sectional and longitudinal dataset.

The proportions in the longitudinal and cross-sectional dataset are slightly different for mathematics and reading, but in general the patterns are quite similar. For phonics the proportions in the cross-sectional and longitudinal data show the largest difference, namely 44% versus 57%. This difference reflects the previously reported lower effect of being in the upper grade in the longitudinal dataset (see table 3). It should be noted, though, that the difference (3.01 – 2.31 = .70) between the effects in both datasets cannot be considered statistically significant as it hardly exceeds the standard errors in either dataset (.61 and .64).

Figure 3



Another matter of concern is whether the analyses of both datasets yield consistent estimates of the grade effect for each school. Using MLwiN the predicted grade effect per school can be obtained. These effects are graphically displayed in figures 4a, 4b and 4c. Appendix 2 provides the exact figures for each school. The effects relate to model 3, i.e. the effect of one year extra schooling after taking into account the linear and quadratic effect of age on achievement.

In general, the figures show a considerable degree of consistency between the longitudinal and cross-sectional data. The grade effect that is estimated for a school in the cross-sectional dataset is usually similar to the effect estimated with the longitudinal data. Nevertheless the effects are never completely identical. The correlation between both effects is modest for phonics (.52) and strong for mathematics and reading (.71 and .78 respectively). The figures 4a, 4b and 4c also reveal that the effect of one year schooling is nearly always positive. Negative effects are exceptional, but they do occur. For one school the grade effect on mathematics achievement is negative in both datasets, whereas none of the schools shows a negative effect for reading. With regard to phonics there is one school with a negative effect, but this is only found in the longitudinal dataset. In the cross-sectional dataset the same school gets a small positive effect.

Figure 4a

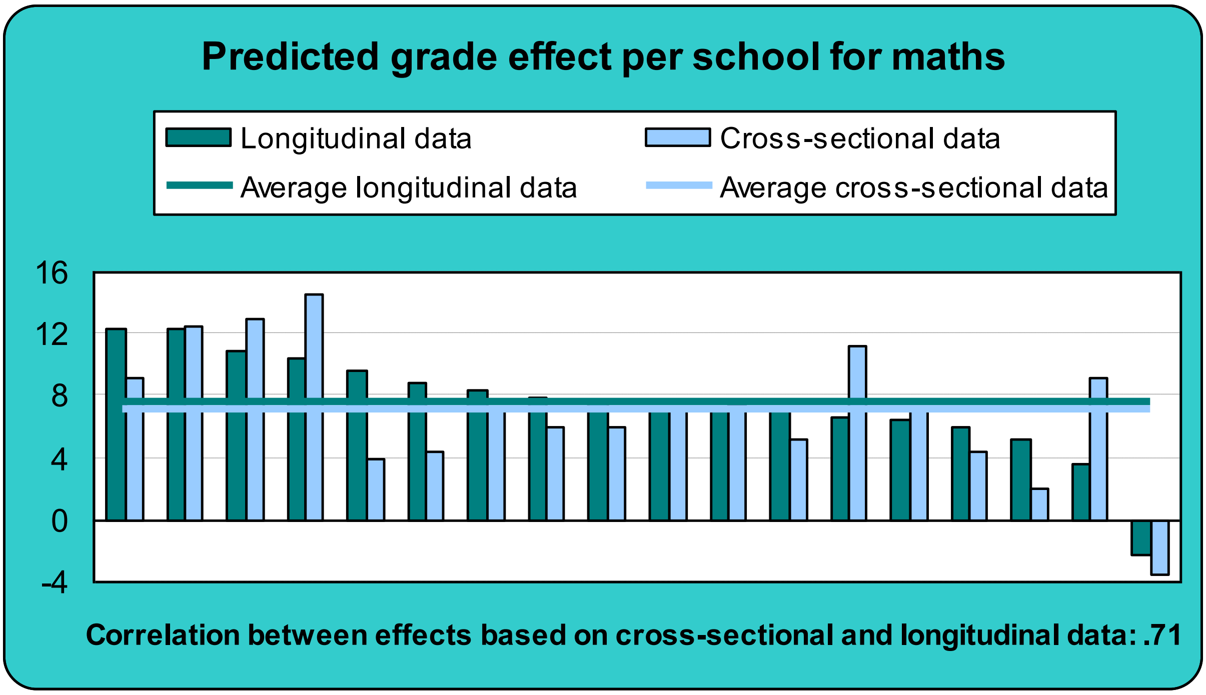


Figure 4b

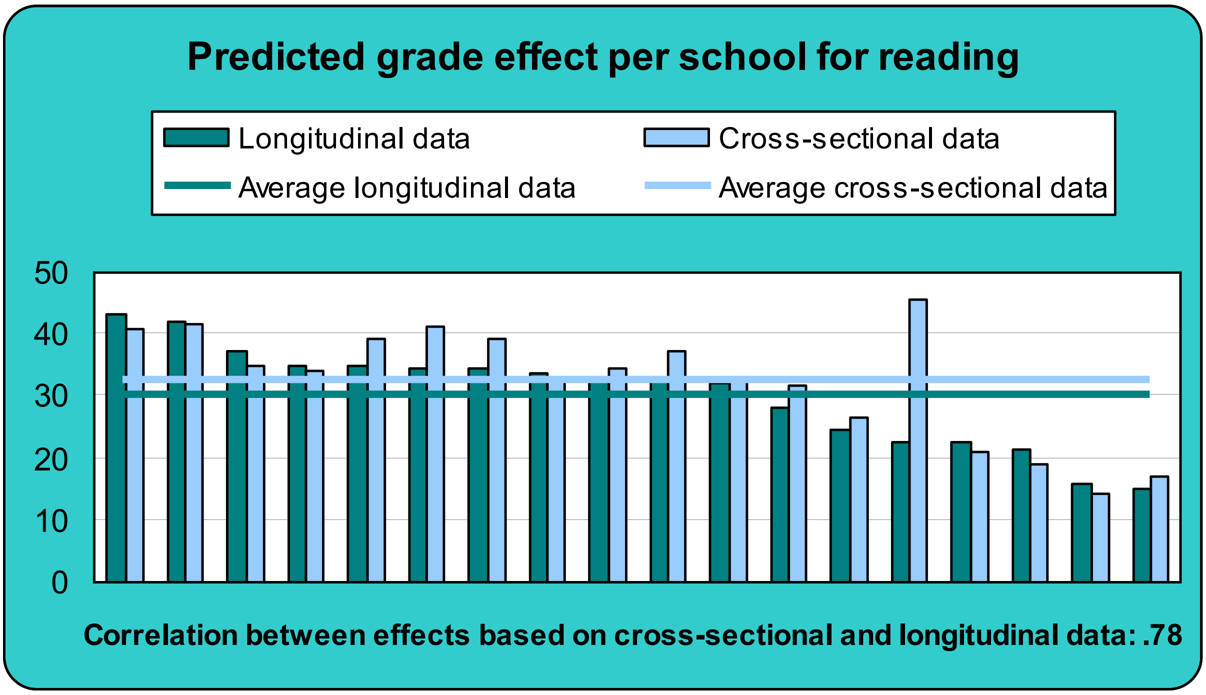
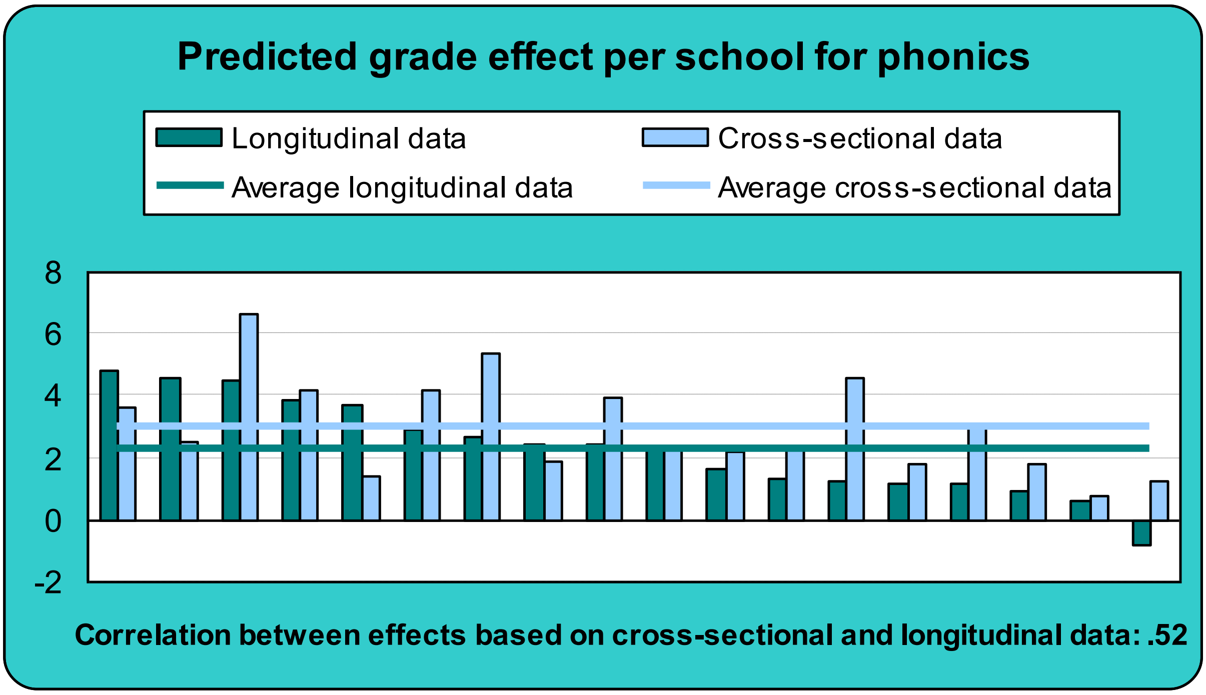


Figure 4c



Summary and discussion

The key question of this study is whether it is possible to make a valid assessment of the effects of schooling using cross-sectional data when making use of the regression-discontinuity approach. Our analyses focused on the effect of one year schooling and its variation across schools. None of the findings for the cross-sectional data appear to differ significantly from the findings produced by the longitudinal data. It should be noted, however, that the sizes of the datasets in this study are modest, so that only fairly large differences will meet moderate significance criteria. Therefore it is advisable to consider the observed differences between the datasets as well. With regard to mathematics and reading the grade effects, i.e. the discontinuities between the oldest pupils in the lower grade and the youngest ones in the upper grade are highly similar for both datasets (see figures 2a and 2b). The percentages of the differences in achievement between both grades that can be accounted for by schooling are only slightly different between the two datasets for mathematics and reading (see figure 3). With regard to phonics the correspondence between the findings for the cross-sectional and longitudinal data is less strong, but still the results are largely similar.

The analyses show more variance of the grade effect across schools for all three subjects in the cross-sectional dataset. Whether cross-sectional data consistently produce higher estimates for the school level variance of the grade effect is a question that requires further study. It seems possible that the variance of the grade effect is somewhat overestimated when the analysis is based on cross-sectional data. In the case of longitudinal data the variation of the grade effect at both the school and pupil level can be assessed, which may lead to a more accurate and possibly lower estimate of the school level variance. Our findings suggest that not only the estimated variances of the grade effects tend to be larger for cross-sectional data, but also the corresponding standard errors. In the present study the consequence is that confidence intervals for the estimates based on cross-sectional data would still include the values estimated with the longitudinal dataset. The predicted grade effects per school reveal a considerable degree of consistency across both datasets. The correlations are moderate for phonics (.52) and strong for reading and mathematics (.78 and .71 respectively).

In most respects the findings of this study suggest an affirmative answer to the question whether it is possible to assess the effect of schooling without controlling for prior achievement. The risk of obtaining biased estimates of the grade effect when the analysis is based on cross-sectional data seems limited. This points to an important practical advantage of the regression-discontinuity approach. Collecting longitudinal data requires more time, effort and money than cross-sectional data. Moreover longitudinal data often suffer from bias due to selective attrition.

In addition to the practical advantages great importance can be attached to the conceptual benefits of the regression-discontinuity approach. First of all, it gives a different perspective on the impact of schooling. Whereas the multilevel analyses that focus on relative differences between schools find that ten to fifteen percent of the variance in test scores is situated at the school level, the regression-discontinuity approach indicates that schooling accounts for a much larger proportion of the cognitive development of pupils. The percentages reported in this study – on average 50% – are relatively low in comparison to those reported by Cahan & Davis (1987) and Luyten (2006). As mentioned in the introduction section these percentages relate to an aspect of the effect of schooling that is different from what is expressed by the usually reported percentages of school level variance. When these percentages are converted to effect sizes that have been defined in relation to interventions in which there is a control and an experimental group, it is found that 10% to15% school level variance corresponds to an effect size of .67 to.70 (see Tymms, 2004). Preliminary calculations on the findings reported in table 3 suggest that converting the effects of one year extra schooling yield similar effect sizes.

Probably the most important advantage of the regression-discontinuity approach is that it allows for an assessment of the absolute effect of schooling. Clearly the usual school effectiveness research “school effects” merely relate to a school’s relative position in comparison to other schools. Thus one will always find that 50% of the schools score above average and 50% below, regardless the overall quality of education in a country. With the regression-discontinuity approach it is possible to express the effect of schooling on a scale that has a meaningful zero-level. A zero effect implies that the difference in test scores between two consecutive grades can be attributed completely to the fact that the pupils in the upper grade are one year older. In a multilevel framework this effect can be estimated per school, so that it is possible to detect schools with a zero or even negative impact on the development of the pupils.

In the present study one school out of eighteen was found to have a negative effect for mathematics in both datasets. Another school showed a negative effect with regard to phonics in the longitudinal dataset, whereas a small positive effect was found in the cross-sectional dataset for this school. The study by Luyten (2006), which also applies the regression-discontinuity approach, reports a negative effect in 23% and 27% of the English primary schools for mathematics and science achievement respectively. These findings came from a secondary analysis of the TIMSS-95 data and relate to pupils in grade three and four in a large number of countries (Mullis et al., 1997; Martin et al., 1997). The relative effects of schooling found in the present study (50% on average) are considerably larger than the percentages reported for the English pupils in the analyses of the TIMSS-95 data, which are 38% for maths and 34% for science. Possible explanations for these differences may be the age of the pupils and the curriculum alignment of the tests used to measure pupil achievement. The TIMSS-95 data relate to nine and ten year olds, while the present analyses focused on four and five year olds. We know that the effect of age diminishes as children grow older (Jones and Jellis, 2005) but that relates to the age/achievement relationship within a year group. We hypothesise that there is an additional increment by grade effect which is greatest for the youngest children. Consider learning to read. This is a major accomplishment often achieved during the first year at school and it is difficult to conceive of a jump of comparable magnitude in later years. Similar examples can be given for mathematics. The youngest children make the greatest leaps. Finding a school where young children have not moved forward is a great surprise. With regard to curriculum alignment we cannot be sure. The test-curriculum overlap in the present study is not clear and it may also not be tight in the case of TIMSS-95.

The findings from this study suggest that there is no need to control for background characteristics in a study that uses the regression-discontinuity approach in order to assess the effect of education. Nevertheless, studies dealing with the effects of background characteristics such as gender, family background and ethnicity on the outcomes of schooling are still valuable. In the context of a regression-discontinuity analysis the research should focus on the question to what extent the effect of one year extra schooling is stronger or weaker for certain groups of pupils. This would require including interaction effects of student characteristics with grade in the data analysis. Effects of school and classroom variables can be modelled in the same manner. See also the contribution by Verachtert et al., which reports different learning rates by socioeconomic background.

In the present study we did not address the impact of student and school characteristics, but it is quite plausible that these variables account to some extent for the variance of the grade effect across schools. What the results of our analyses show, is the average effect of one year schooling across different kinds of pupils and schools. Still it seems likely that, for example, the grade effect is stronger for pupils with highly educated parents. In that case, the educational level of the parents can account for part of the variance of the grade effect between schools.

As has already been noted the present study relates to a fairly small dataset and to a specific group of pupils, namely four and five year olds, but we expect that future research will provide further support for the assumption that in many respects cross-sectional data suffice for a valid assessment of the impact of schooling. This requires replications of the present study on larger datasets and other age ranges. Another important topic for subsequent studies is the comparison of the interaction effects of pupil and school characteristics with the grade effect between longitudinal and cross-sectional data.

In conclusion we want to emphasize the basic requirements for application of the regression-discontinuity approach in future research on educational effectiveness. First of all the approach can only be applied if the data relate to two (or more) *adjacent* grades. In their contribution Kyriakides & Luyten illustrate how the approach can be applied to a dataset that covers more than two grades in a row. However, regression-discontinuity can not be applied to assess the effect of schooling if the dataset relates to non-neighboring grades (e.g. grades 2 and 4).

Another crucial requirement involves the comparability of the outcomes measures across grades. In the present study the same test is used in both grades. This may not always be feasible, especially if the analysis covers a wide range of grades. Using exactly the same tests in all grades is not necessary. The key point is that the tests used in the various grades relate to a common scale. This can be achieved by equating techniques based on classical test theory or item response theory.

Correct modeling of the age-achievement relation is another point of crucial importance. If a linear function is fitted, while the real relation is curvilinear, the estimated effect of schooling may be biased. It is also worth mentioning that the approach does not require a strong or statistically significant relationship between age and achievement. As students grow older one may expect a diminishing link to age, but this does not affect the validity of the approach. If the relation between the selection criterion and the outcome measure were close to zero, one would actually approximate random assignment to grades (Trochim, 1984). However, it is of the utmost importance to emphasize that regression-discontinuity assumes a strict adherence to the cut-off point. In the present study the percentage of students with non-standard school careers is negligibly small (1.5%), but the 'misclassification' of pupils presents a serious problem regarding the usefulness of the regression-discontinuity approach in many countries where other factors besides date of birth influence assignment to grades. Grade retention is undoubtedly one of the main reasons why assignment to grade level does not always agree with a cut-off point based on date of birth. Only in a restricted number of the countries is the proportion of 'misclassifications' small enough to obtain reliable estimates of the grade effect without considering other factors. The greatest challenge will be to develop valid methods for taking into account the effects of these other factors when assessing the effect of schooling. Addressing the 'misclassification' issue as a selection bias problem seems a promising approach. When trying to assess the grade effect we are faced with the problem that assignment to grades (in the guise of grade-repeating or grade-skipping) is partly determined by factors that affect the outcomes of learning (Luyten & Veldkamp, 2008).

References

Adams, S., Alexander, E., Drummond, M. J., & Moyles, J. (2004). *Inside the Foundation Stage: Recreating the reception year: Final Report* (No. PR18). London: Association of Teachers and Lecturers.

Cahan, S. & Davis, D. (1987). A between-grades-level approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal, 24*(1), 1-12.

Jones, P.G. & Jellis, C. (2005), Age Effects within the Primary School Assessments, paper given at International Association for Cognitive Education and Psychology, Durham 2005.

Judd, C.M. & Kennedy, D.A. (1981). *Estimating the effects of social interventions.* New York: Cambridge University Press.

Hill, P. & Rowe, K.J. (1996). Multilevel Modelling in School Effectiveness Research. *School Effectiveness and School Improvement, 7*(1), 1-34.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling, regression discontinuity applied to TIMSS-95. *Oxford Review of Education, 32*(3), 397-427.

Luyten, H. & Veldkamp, B. (2008). *Assessing the effect of schooling with cross-sectional data, between grades differences addressed as a selection-bias problem,* Paper presented to the annual conference of the Dutch Educational Research Association, Eindhoven.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzales, Smith, T.A. & Kelly, D. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill: TIMSS international study center, Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzales, Kelly, D. & Smith, T.A. (1997). *Mathematics achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill: TIMSS international study center, Boston College.

Opdenakker, M-C. & Van Damme, J. (2000). Effects of Schools, Teaching Staff and Classes on Achievement and Well-Being in Secondary Education: Similarities and Differences between School Outcomes. *School Effectiveness and School Improvement,11*(2), 165-196.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., Lewis, T. (2000). *A user's guide to Mlwin.* London: Multilevel Models Project, Institute of Education, University of London.

Scheerens, J. & Bosker, R.J. (1997). *The foundations of educational effectiveness.* Oxford: Pergamon.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston/New York: Houghton Mifflin Company.

Trochim, W.M.K. (1984). *Research design for program evaluation, the regression-discontinuity approach.* London: SAGE Publications.

Tymms, P. (1999a). *Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators*. London: David Fulton Publishers.

Tymms, P. B. (1999b). Baseline assessment, value-added and the prediction of reading. *Journal of Research in Reading, 22*(1), 27-36.

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean?* (pp. 55-66). Slough: National Foundation for Educational Research.

Tymms, P., & Albone, S. (2002). Performance Indicators in Primary Schools. In A. J. Visscher & R. Coe (Eds.), *School Improvement Through Performance Feedback* (pp. 191-218). Lisse/Abingdon/Exton PA/Tokyo: Swets & Zeitlinger.

Tymms, P., & Coe, R. (2003). Celebration of the Success of Distributed Research with Schools: the CEM Centre, Durham. *British Educational Research Journal, 29*(5), 639-653.

Appendix 1: Random effects in models 1 and 2 (compare table 3)

	MATHEMATICS				READING				PHONICS			
	Cross-sectional data		Longitudinal data		Cross-sectional data		Longitudinal data		Cross-sectional data		Longitudinal data	
MODEL 1												
School level												
Variance intercept	9.74	(4.19)	12.48	(5.13)	44.70	(19.08)	103.86	(38.08)	4.21	(1.62)	5.18	(2.00)
Variance grade effect	21.14	(8.89)	13.17	(4.96)	156.50	(71.81)	102.74	(43.70)	3.23	(1.42)	2.59	(1.04)
Covariance	-6.14	(4.81)	-4.08	(3.76)	8.91	(26.27)	3.85	(29.00)	-3.25	(1.41)	-3.39	(1.37)
Correlation	-.43(not sign.)		-.32(not sign.)		.11(not sign.)		.04(not sign.)		-.88		-.92	
Pupil level												
Variance intercept	75.59	(4.46)	65.96	(3.87)	332.46	(19.60)	206.56	(12.12)	15.95	(.94)	17.75	(1.04)
Variance grade effect	---	---	34.99	(2.05)	---	---	656.15	(38.47)	---	---	12.57	(.74)
Covariance	-9.88	(2.77)	-22.51	(2.20)	400.08	(34.62)	135.44	(16.27)	-2.63	(.57)	-9.81	(.74)
Correlation	cannot be computed		-.47		cannot be computed		.37		cannot be computed		-.66	
MODEL 2												
School level												
Variance intercept	10.82	(4.45)	11.27	(4.63)	49.93	(20.70)	102.41	(37.22)	4.35	(1.66)	5.43	(2.09)
Variance grade effect	22.51	(9.21)	13.15	(4.95)	136.50	(63.73)	103.01	(43.78)	3.14	(1.38)	2.64	(1.07)
Covariance	-7.61	5.13	-3.64	(3.55)	8.47	(25.80)	-3.14	(28.62)	-3.31	(1.42)	-3.56	(1.43)
Correlation	-.49(not sign.)		-.50(not sign.)		.10(not sign.)		-.03(not sign.)		-.90		-.94	
Pupil level												
Variance intercept	66.70	3.94	59.64	(3.50)	318.29	(18.76)	181.54	(10.65)	15.42	(.91)	16.66	(.98)
Variance grade effect	---	---	34.99	(2.05)	---	---	656.15	(38.47)	---	---	12.56	(.74)
Covariance	-7.34	2.50	-21.21	(2.09)	368.97	(32.36)	115.06	(15.09)	-2.51	(.55)	-9.41	(.72)
Correlation	cannot be computed		-.46		cannot be computed		.33		cannot be computed		-.65	

The figures in between brackets denote the standard errors.
Model 1 includes only the effect of grade (fixed and random).
Model 2 includes the effect of grade (fixed and random) and a linear age effect (fixed).
See table 3 for the fixed effects of grade and age in models 1 and 2.

Appendix 2: Predicted grade effects per school

School	Number of pupils		Effect for mathematics		Effect for Reading		Effect for phonics	
	Lower grade	Upper grade	Cross-sectional	Longi-tudinal	Cross-sectional	Longi-tudinal	Cross-sectional	Longi-tudinal
1	63	46	14.63	10.42	37.16	32.89	3.96	2.43
2	12	18	4.49	8.87	32.56	32.30	3.66	4.79
3	53	45	7.34	7.21	41.01	43.11	4.20	2.94
4	17	13	7.40	7.12	34.13	35.02	3.02	1.14
5	5	7	5.97	7.77	39.15	34.86	1.44	3.72
6	24	27	-3.49	-2.20	41.75	42.16	2.56	4.57
7	23	8	3.99	9.64	35.03	37.28	0.77	0.59
8	30	41	7.14	8.33	34.38	33.12	5.40	2.72
9	42	38	6.02	7.91	26.57	24.66	2.33	1.34
10	56	59	2.10	5.28	14.19	15.68	2.24	1.64
11	15	25	9.14	12.38	33.11	33.64	2.39	2.29
12	18	4	11.30	6.66	31.91	28.07	4.59	1.25
13	60	55	12.99	10.95	41.08	34.46	6.67	4.52
14	30	30	4.47	5.97	17.21	14.95	4.18	3.87
15	25	9	9.11	3.65	45.57	22.81	1.28	-0.81
16	37	37	12.46	12.31	39.48	34.42	1.84	1.18
17	21	53	7.15	6.52	21.18	22.62	1.78	0.92
18	62	84	5.23	7.10	19.18	21.37	1.87	2.47